Damodara sree Ranganath Yadav* et al.
(IJITR) INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND RESEARCH
Volume No.4, Issue No.6, October – November 2016, 4959-4961.

A Technical Report On Big Data, Tools And Technologies

DAMODARA SREERANGANATH YADAV
Technical lead
Epsilon India.

Abstract: **As the data is growing day by day and managing that data is becoming a difficult task. To overcome this problem, Big data is used. Big data consists of huge data sets and it will have complex data structures to store them. Big data contains the data like social media data, companies information etc. The existing tools are unable to manage this type of data and extracting information which is needed by the user is becoming a difficult task. If the company wants to know some information and based on that if they want to make a decision it is not happening with the existing tools. All the analysis and extracting information is done by the big data analytics tools. This paper concentrates on the importance of the big data, big data analytics and the tools which will be applied on the big data.**

Keywords: **Big Data; Big Data Analytics; Analysis;**

I. INTRODUCTION

Main growth of the organization will depends on the data they store. This data will be useful for them in later phases to know what was happened in the past. Based on the available information they can have the analysis and takes a decision. If organization is functioning without any storage of data then it will loose all these type of capabilities. Data may be of different types like regular transactions which are had in the organizations or the personal information of the customers or any reviews on the organization or on the products which are developed by the customers. To store all this inter related data and to provide the information needed by the user, existing data storage tools are not sufficient. As technology is growing so many tools are coming into the market to meet the current requirements of the customer. From that Big data is one among them which stores huge data and applies the analysis on it and gives the needed information to the user. The data which is storing is not constant and it may change dynamically. For these changing environments and to improve the decision making capabilities big data analytics is used. The main contribution of this paper is to analyze the existing literature on big data analytics, tools used by the big data, methods and concepts applied etc.

Figure 1. presents the architecture of the big data analytics. This architecture is divided into 3 layers like infrastructure layer which concentrates on the devices which are used to implement the big data technology, Second layer is computing layer which contains storage tools and programming models in storing the data, third layer is the application layer which provides a kind of communication in query processing, in classifying the data, in having some recommendation with the help of different concepts applied.

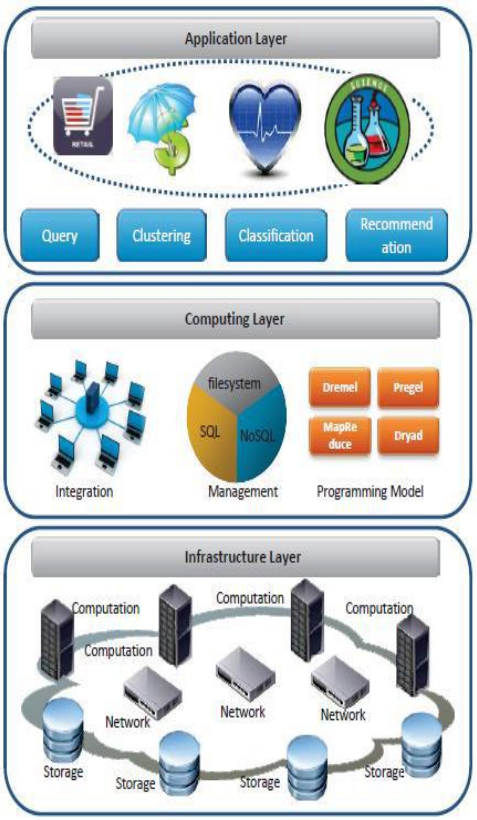

Figure 1: Architecture of the Big data.

Big data is having huge data storage capacity. Generally the existing tools will store megabytes or gigabytes of the data. But big data can able to store the petabytes of the data. By this we can understand big data stores more volume of data. Not only storing huge quantity but also concentrates on more storage repositories like structured, semi-structure and un structured etc. Another important characteristic of the big data is in processing speed. It gives immediate results requested by the user.

II. BIG DATA ANALYTICS

Traditional storage systems are not managing huge data. Data to be stored is increasing day by day. With these requirementsit is becoming a problem to work with traditional data storage systems. All the traditional storage systems is having a limited structures to work with. Unlike traditional storage tools, big data stores terabytes and petabytes of data. In the current generation all the enterprises are storing their data by using the big data because there exists so many tools in analyzing the information existed and gives the required result. The information which is had as a result of analytics may not be known to the user.

While processing the information using big data there will be some difficulties that has to overcome by the big data. One of the main problems is data heterogeneity, when we deal with data personally then we can understand that whatever the data is, but here there will be different types of data will be there with different structure. It is the responsibility of the big data storage to manage all such kind of data. Generally, data will be increasing day by day, but the traditional storage and accessing tools does not have the capacity to have the scalability with storage structures. Another ne is to maintain the speed processing of data. If the required result is not given in the said time interval then there is no use of storing such data. Security and privacy is one of the most important feature which is achieved by the big data management.

III. TECHNOLOGIES TO HANDLE BIG DATA.

One of the tool which is used to handle large data sets is known as Hadoop. This tool is developed by the Google. The components of the Hadoop consists of Hadoop Kernel, MapReduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper.

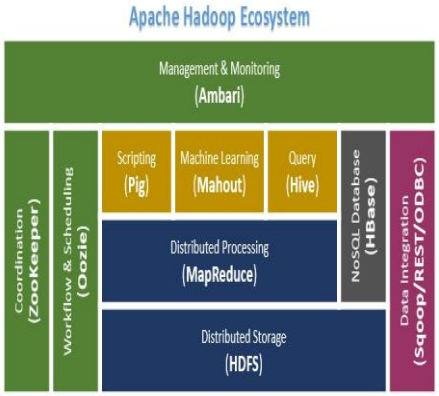

Figure 2: Components of Hadoop.

Figure 2 represents the components of Hadoop system. One of the main component in the Hadoop system Hadoop distributed file system(HDFS). This component stores huge amount of data and it is having auto increment capacity regarding storage. It even survives when there exist any failures in the storage system. There won't be any loss of data. This feature is achieved by maintaining clusters of machines and establishing the coordination among them. If one of the machines fails then the rest of the machines will be taking care of the work.HDFS manages storage on the cluster by breaking incoming files into pieces, called "blocks," and storing each of the blocks redundantly across the pool of servers.

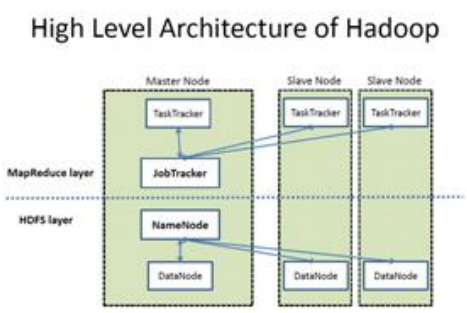

Figure 3: Hadoop Architecture.

HDFS stores all the copies and stores each copy into three different servers.

Coming to the MapReduce frameworkk architecture, this is the processing component of the Hadoop system. This framework implements the operation by serating the data and the operation and both of them will be parallely run to achieve the result set. This is based on the ETL concept which is known as Extraction, Transformation and Loading. All these functioanlities are written as mapreduce jobs in the architecture. The output of these jobs are again stored back into the HDFS files system or any other traditional storage structures.The advantage of map reduce is a large variety ofproblems are easily expressible as Map reducecomputations and cluster of machines handle thousandsof nodes and fault-tolerance. The main disadvantage of this framework is map reduce is Real-timeprocessing, not always very easy to implement,shuffling of data, batch processing.

Components of Map Reduce:

1. **Name Node**: manages HDFS metadata, doesn'tdeal with files directly.
2. **Data Node**: stores blocks of HDFS—defaults replication level for each block: 3.
3. **Job Tracker**: schedules, allocates and monitorsjob execution on slaves—Task Trackers.
4. **Task Tracker**: runs Map Reduce operations.

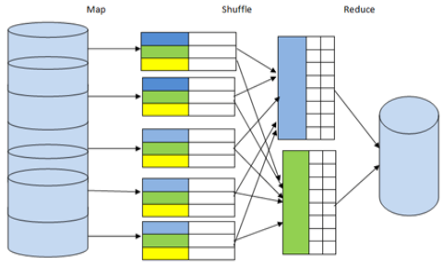

Figure 3: Map Reduce Architecture.

Big data analytics will be used in different fields to provide the value of information to the users of big data. This is used provide the information after implementing the analysis and helps in making a decision. Apart from this it also helps in improvising the customer intelligence and brings the benefits into different industries likeretail, banking, and telecommunications.

It is also used in the supply chain management to know the demand of the market and accordingly it has to inform the manufacturing department to manufacture and at the end it has to supply as per the demand. Performance management is another key area where big data is useful to monitor the performance of the employees. Industries like health care and the government organization will benefit with the implementation of Big data. This can also be used in other areas like maintaining the quality, Risk management and fraud detection etc.

IV. SCOPE OF RESEARCH.

There exist different problems in implementing the big data. If the user is storing only homogenous data then processing the homogenous data is not complex. But every system or user or application uses heterogonous structures to store all types of data. Here we can concentrate more on parallel processing concepts, using different algorithms for streamlining of data and even we can think about in improving the processing speed

V. CONCLUSION

As I have gone through with the introduction of the big data and the uses of it. Then I had presented the importance of the big data analytics, later on I had gone through with the technologies and tools used to implement the big data. At the end I understood that there exists some gap in implementing the concepts to improvise the performance. To fill the gaps one has to concentrate the drawbacks of existing technology and tools, to implement the new algorithms, concepts and in improvising the processing speed etc.

VI. REFERENCES

[1] Bakshi, K.: Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7 (2012).

[2] Sabia and Love Arora: Technologies to Handle Big Data: A Survey.

[3] Hadoop. http://hadoop.apache.org/

[4] Shilpa, ManjitKaur: BIG Data and Methodology-A review, International Journal of Advanced Research in Computer Science and Software Engineering,Volume 3, Issue 10, October 2013

[5] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar: A Review paper on Big data and Hadoop, International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014

[6] Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills: New AnalysisPractices for Big Data. Proceedings of the ACM VLDB Endowment 2(2), 1481–1492(2009)

[7] Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data:The Big Data Revolution! In: Proceedings of the ACM International Workshop on DataWarehousing and OLAP, pp. 101–104 (2011)

[8] Yuri Demchenko "The Big Data Architecture Framework(BDAF)" Outcome of the Brainstorming Session at theUniversity of Amsterdam 17 July 2013.

[9] Tekiner F. and Keane J.A., Systems, Man and Cybernetics(SMC), "Big Data Framework" 2013 IEEE International Conference on 13–16 Oct. 2013, 1494–1499.

[10] Herodotou, H. Lim, G. Luo, N. Borisov, L. Dong, F. B. Cetin, and S. Babu. Starfish: A Self tuning System for Big Data Analytics. In CIDR, pages 261–272, 2011.

